The Cross-Depiction Problem: Recognition and Synthesis of Photographs and Artwork

Peter Hall
Department of Computer Science
University of Bath, UK
pmh@bath.ac.uk

Hongping Cai
Department of Computer Science
University of Bath, UK
H.Cai@bath.ac.uk

Qi Wu
School of Computer Science
University of Adelaide, Australia
qi.wu01@adelaide.edu.au

Tadeo Corradi
Department of Computer Science
University of Bath, UK
ma1tmc@bath.ac.uk

Abstract

The *cross-depiction* is the recognition – and synthesis – of objects whether they are photographed, painted, drawn, *etc.* It is a significant yet under-researched problem. Emulating the remarkable human ability to recognise and depict objects in an astonishingly wide variety of *depictive forms* is likely to advance both the foundations and the applications of Computer Vision.

In this paper we motivate the cross-depiction problem, explain why it is difficult, discuss some current approaches. Our main conclusions are (i) appearance-based recognition systems tend to be over-fitted to one depiction, (ii) models that explicitly encode spatial relations between parts are more robust, and (iii) recognition and non-photorealistic synthesis are related tasks.

Keywords: cross-depiction, classification, synthesis, feature, spatial layout, connectivity, representation.

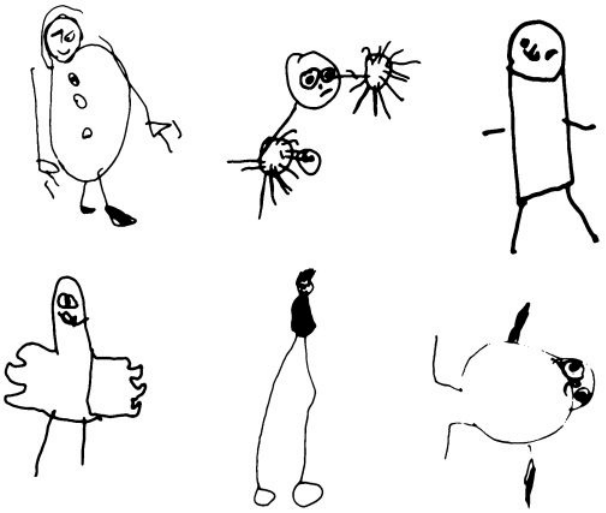

Figure 1: Children's drawings.

1 Introduction

Many years ago, I took my young children to the zoo. I showed them a simple drawing of a giraffe; bright coloured areas, black lines. When the children got to the zoo, they had no problem at all identifying the giraffe, or the camel, the lion, *etc.* What is more, they could make recognisable depictions of these animals.

The children were exhibiting (at least) two abilities. One is to generalise from a specific instance to a class, and the other is to generalise from a depiction (in that case, a particular style of artwork) to real life. Children generalise equally well across depictions; they would have recognised photographs of the animals equally well. Humans are able to recognise objects in an astonishing variety of forms. Whether photographed, drawn, painted, carved in wood, people can recognise horses, bicycles, people, *etc.* Furthermore, the ability to draw and paint – even from memory – is a strong indicator that in humans at least, recognition and synthesis are related.

The ability of humans to recognise regardless of depiction is such a everyday occurrence that it can often pass without being noticed. Yet it is an astonishing ability that cannot be matched by any current algorithm. Even the very best recognition algorithms – including deep learning – fail to cope with the cross depiction problem. Indeed, all algorithms we have empirically tested exhibit the same general behaviour: all show a significant drop in performance when presented with an inhomogeneous data set, and fall further still when trying to recognise a drawn object after being trained only on photographic examples. Some algorithms are more pronounced then others in this trend – those that explicitly encode spatial relations tend to be more robust.

The inability of all contemporary approaches to cope with the cross-depiction problem is a significant literature gap. Cross-depiction forces one to consider which visual attributes are necessary for recognition, and which are merely sufficient. That is, one may sensibly ask: *Which properties of an object class are in-*

variant (or close to invariant) given over variations in depictive style?. The specific appearance among different depictive styles varies to a much greater degree than that due to lighting changes, but still people can recognise them. Children's drawings, as in Figure 1, are both highly abstract and highly variable, yet contain sufficient information for objects to be recognised by humans, but not computers. Equally overlooked is the fact that no computer is yet able to draw as a child.

Learning the specifics of each depiction seems at best unappealing, not least because the gamut of possible depictions is potentially unlimited. Rather, the question is: *What abstraction do these classes have in common that allow them to be recognised regardless of depiction?* It is a unavoidable questions that push at the foundations of Computer Vision.

A machine that is able to recognise regardless of depiction would provide a significant boost to current applications, such as image search and rendering. For example, given a photograph of the Queen of England, a search should return all portraits of her, including postage stamps that capture her likeness in bas-relief. Searching heterogeneous data sets is a real problem for the creative industries, because they archive vast quantities of material in a huge variety of depictions – a problem that requires visual class models to span depictive styles. Non-photorealistic rendering from images and video would be boosted too, not least because highly aesthetic renderings depend critically on the level of abstraction available to algorithms. Picture making is nothing like tracing over photographs: humans draw what they know of an object, not what they see – computers should do like wise.

One of our guiding principles has been that the cross-depiction problem acts to unite the synthesis and analysis of images. The rationale is that people find it at best very difficult to draw objects they cannot recognise; more exactly, people tend to draw objects they can see in a manner that if highly influenced by what they know of them. This is most obvious in children, who draw the sky at the top of their pictures, and eyes at the top of heads, often they will draw cars with four wheels, and so on. But it is evident in the artwork of adults too. Indeed, students at Western Art schools are given extensive life-drawing classes with the exact purpose of teaching them to draw what is seen rather than what is known. For example, early students often draw the hands, feet, and faces in proportion to direct measures rather than as seen when foreshortened: the students' knowledge allows them to compensate for perspective effects.

The key for computational emulation of the human ability is, we argue, *representation*. It is reasonable to seek a single representation that supports both the recognition and the synthesis of objects.Even so, from an "engineering" point of view the problems of recognition and synthesis seem sufficiently far apart that different representation are needed. Therefore, we will consider representations that are suitable for each, and then conjecture as to what a single representation might look like.

In summary then, there are two important reasons to study the cross-depiction:

1. The 'foundational' problem: We are forced to think very carefully about how to model object classes.

2. The 'practical' consequences: Solving the cross-depiction problem will open many robust applications in web search, computer graphics, and other areas.

This paper establishes there is a literature gap, it shows that feature based approaches alone are not sufficient for cross-depiction, and that representations that take connectivity and spatial layout into account perform better. It suggests future avenues in terms of object class representation. As a note: in this paper, we use the term *photograph* as a short hand for "natural image", and the term *artwork* as all other images.

2 Related Literature

The Computer Vision literature distinguishes between *classification* (does this image contain an object of class X, or not?) and *detection* (an object of class X is at this place in this image). Yet lay language makes no sharp distinction; we use the term *recognition* to mean both classification and detection, which is closer to lay usage.

The is a vast literature in Computer Vision to address recognition. Yet almost no prior art addresses the cross-depiction, which is surprising given its genuine potential for advancing Computer Vision both in its foundations and in its applications.

Of the many approaches to visual object classification, the bag-of-words (BoW) family [14, 40, 48] is amongst the most widespread. It models visual object classes as histograms of visual words; these words being clusters in feature space. Although the BoW methods address many difficult issues, they tend to generalise poorly across depictive styles (see Sec. 3). Alternative low-level features such as edgelets [29, 57] may be considered, or mid-level features such as region shapes [36, 31]. These features offer a little more robustness, but only if the silhouette shape is constrained – and only if the picture offers discernible edges, which is not the case for many artistic pictures (Turner's paintings, for example).

Deformable models of various types are widely used to model object classes for detection tasks, including several kinds of deformable template models [8, 9] and a variety of part-based models [1, 10, 20, 19, 21, 25, 41]. In the constellation models from [21], parts are constrained to be in a sparse set of locations, and their geometric arrangement is captured by a Gaussian distribution. In contrast, pictorial structure models [20, 19, 25] define a matching problem where parts have an individual match cost in a dense set of locations, and their geometric arrangement is captured by a set of spring connecting pairs of parts. In those methods, the

Deformable Part-based Model (DPM) [19], is widely used. It describes an object detection system based on mixtures of multi-scale deformable part models plus a root model. By modelling objects from different views with distinct models, it is able to cope with large variations in pose. None of these directly address the cross-depiction problem.

Shape has also been considered. Leordeanu *et al.* [42] encode relations between all pairs of edgels of shape to go beyond individual edgels. Similarly, Elidan *et al.* [16] use pairwise spatial relations between landmark points. Ferrari et al. [23] propose a family of scale invariant local shape features formed by short chains of connected contour segments. Shape skeletons are the dual of shape boundary, and also have been used as a descriptor. For example, Rom and Medioni [47] suggest a hierarchical approach for shape description, combining local and global information, to obtain skeleton of shape. Sundar *et al.* [55] use skeletal graph to represent shape and use graph matching techniques to match and compare skeletons. Shock graph [52] is derived from skeleton models of shapes, and focus on the properties of the surrounding shape. Shock graphs are obtained as a combination of singularities that arise during the evolution of a grassfire transform on any given shape. In particular, the set of singularities consists of corners, lines, bridges and other similar features. Shock graphs are then organised into shock trees to provide a rich description of the shape.

Algorithms usually assume that the training and test data are drawn from the same distribution. This assumption may be breached in real-world applications, leading to domain-adaptation methods such as transfer component analysis (TCA) [45], which transfer components from one domain to another. Both sampling geodesic flow (SGF) [30] and geodesic flow kernel (GFK) [29] use intermediate subspaces on the geodesic flow connecting the source and target domain. GFK represents state-of-the-art performance on the standard cross-domain dataset [22]; it has been used to classify photographs acquired under different environmental conditions, at different times, or from different viewpoints.

Cross-depiction problems are comparatively less well explored. Some work is very specific – Crowley and Zisserman take a weakly supervised approach, using a DPM to learn figurative art on Greek vases [13]. Others develop the problem of searching a database of photographs based on a sketch query; edge-based HOG was explored in [34], Li *et* al. [43]. Other have investigated sketch based retrieval of video [35, 7].

Approaches to the more general cross depiction problem are rare. Matching visually similar images has been addressed using self similarity descriptors [50]. It relies on a spatial map built from correlations of small patches; it therefore encodes a spatial distribution, but tends to be limited to small rigid objects. Crowley and Zisserman [12] provide the only example of domain adaptation we know of specifically designed for the cross depiction problem; they train on photographs and then use midlevel patches to learn spatial consistencies (scale and translation) that allow matching from photographs into artwork. Their method performs well in retrieval tasks for 11 object classes in databases of paintings.

Classification, rather than matching, has also been studied. Shrivista *et al* [51] show that an Exemplar SVM trained on a huge database is capable of classification of both photographs and artwork. A less computationally intensive approach has been proposed [62] using a hierarchical graph model to obtain a coarse-to-fine arrangement of parts with nodes labelled by qualitative shape [60]. Wu *et al* address the cross-depiction problem using a deformable model [59]; they use a fully connected graph with learned weights on nodes (the importance of a nodes to discriminative classification), on edges (by analogy, the stiffness of a spring connecting parts), and multiple node labels (to account to different depictions); a method tested on 50 categories. Others use no labels at all, but rely on connection structure alone [2] or distances between low-level parts [42].

Deep learning has recently emerged as a truly significant development in Computer Vision. It has been successful on conventional databases, and over a wide range of tasks, with recognition rates in excess of 90%. Deep learning has been used for the cross-depiction problem, but its success is less clear cut. Crowley and Zisserman [11] are able to retrieve paintings in 10 classes at a success rate that does not rise above 55%; their classes do not include people. Ginosar *et al* [26] use deep learning for detecting people in Picasso paintings, achieving rates of about 10%.

Other than this paper, we know of only two studies assessing the performance of well established methods on the cross depiction problem. Crowley and Zisserman [12] use a subset of the 'Your Paintings' dataset [3], the subset decided by those that have been tagged with VOC categories [17]. Using 11 classes, and objects that can only scale and translate, they report an overall drop in per class Prec@k (at $k = 5$) from 0.98 when trained and tested on paintings alone, to 0.66 when trained on photographs and tested on paintings. Hu and Collomosse [34] use 33 shape categories in Flickr to compare a range of descriptors SIFT, multi-resolution HOG, Self Similarity, Shape Context, Structure Tensor, and (their contribution) Gradient Field HOG. They test a collection of 8 distance measures, reporting low mean average precision rates in all cases.

Regarding synthesis, non-photorealistic rendering from photographs is germane to our paper. Almost all of the NPR from photographs literature concerns the development of image filtering of one kind or another, see for example [39] for a review. However, such algorithms fail to emulate the process of human produced arts, which is inevitably about *abstraction* of some kind, meaning a summary of the object or scene being drawn. Moreover, humans can and do draw (and paint) from memory.

3 Representations for Recognition

Here we will consider representations for recognition of object classes, regardless of how they are depicted. We describe representations we have used, and benchmark some of them against datasets we have created.

3.1 Feature Based Representations

As already mentioned in Section 2, Bag of Word (BoW) models for object classes is widespread. BoW models are premised on the assumption that object classes can be distinguished from the relative proportion of discriminative image patches in an un-ordered collection. Since "words" in the context of images means an image patch, the consequence of the this assumption is that words in patch must exhibit low variation – they must be similar.

Intuitively, this "BoW assumption" is violated when the datasets contain both photographs and artwork; our intuition is confirmed by experiments. In order to see how the local features affect the performance in cross-depiction classification, we test a range of different features, *e.g.*, ***SIFT*** [44], ***Geometric Blur*** (GB) [4],***Self-similarity desciptors*** (SSD) [5], ***Histogram of Oriented Gradient*** (HOG) [15], and ***Edge-based HOG*** (eHOG) [33].

The BoW we use is the spatial pyramid [**?**], as it is well known and widely used. Given a set of labelled training images, local descriptors are computed on a regular grid with multiple-sized regions. A vocabulary of words is constructed by vector quantisation of local descriptors with k-means clustering ($k = 1000$). To construct a visual class model (VCM) each image is partitioned into L levels of increasingly fine cells ($L = 2$ in our experiments). A histogram of word occurrences is built for each cell; concatenating these histograms encodes the image with a 5000 dimensional vector. A one-versus-all linear SVM classifier is trained on a χ^2-homogeneous kernel map [58] of all training histograms. Given a test image, the local features are extracted in the same way as in the training stage, mapped onto the codebook to build a multi-resolution histogram, which is then classified with the trained SVM.

We evaluate the algorithms on Photo-Art-50 dataset [59] which contains 50 distinct object classes (see Figure 2), with between 90 and 138 images for each class. Each class is approximately half photographs and half artwork. All 50 classes appear in Caltech-256; a few also appear in PASCAL VOC Challenge [17] and ETH-Shape dataset [24].

As can be seen in Table 1, none of the BoW methods performs well in recognition over a heterogeneous database such as ours. We also used ***Fisher Vectors (FV)*** [46], which instead describe the distribution of statistics of local features inside each cluster. Consistent with the observation in [46], it outperforms BoW-SIFT by 2-3% in all 'train-test' settings. In spite of such an improvement, FV still suffers from significant performance drop in the condition of different training and test depiction domains.

In summary, all methods exhibit comparably high performance with homogeneous data comply with the "low variation" assumption (good for photographs) but show a fall when faced with heterogeneous data (photographs and artwork). The fall is most distinct when BoW and Fischer Vectors are trained on photographs and tested on artwork – suggesting the representation is over-fitted to photographic data. Due to the very different distribution of photo and art domains, it is natual to resort to the domain adaptation techniques. In the following section, we will investigate how well the domain adaptation could bridge the gap.

3.1.1 Domain Adaptation

Domain adaptation is a process by which a representation built initially for one domain is allowed to somehow adapt to cover a second. Some may say that photographs and artwork belong to different domains, so that domain adaptation may overcome the problems we see with BoW and Fischer Vectors.

Excellent domain adaptive methods include, but are not limited to [30, 29, 22, 49, 28]. They show clear benefits for photographs captured under different conditions. We tested some of these (details below) using photographs as a source domain for the initial model, which we adapted to the target domain of artwork. Recall Table 1 shows this case to be the most difficult for BoW and Fischer Vectors. We also tested adaptation in the reverse direction (from art to photographs, still difficult for BoW and FV).

Specifically, we implemented and tested two variants of ***Geodesic Flow Kernel***(GFK) [29]: GFK_PCA projects original features in both domains (source photogrpah and target artwork) onto a 49 dimensional subspace via with PCA; GFK_LDA uses supervised dimensionality reduction via linear discriminant analysis – on the source domain only. ***Subspace Alignment*** (SA) [22] project $\mathcal{S}$ and $\mathcal{T}$ to respective subspaces. Then, a linear transformation function is learned to align the two domains.

The results for these three methods are shown in Table 1. They suggests that domain adaptation using feature representations are not effective.

3.2 Models with Spatial and Structural Information

As Table 1 shows, feature based representations are poorly suited to the task of recognition in the cross-domain problem; even domain adaptation proves ineffective. This section describes representations that take spatial and structural relations into account.

3.2.1 Structure and Shape

We have used structure alone as a representation[2]. Each class representation was a spatially weighted

Figure 2: Top: Photo-Art-50 dataset [59]: containing 50 object categories. Each category is displayed with one art image and one photo image.

model		BoW					FV	GFK_PCA	GFK_LDA	SA
train	test	SIFT	GB	SSD	HOG	eHOG	SIFT	SIFT	SIFT	SIFT
P	P	84	77	66	72	70	87	-	-	-
M	P	80	72	58	65	63	84	-	-	-
A	P	64	60	39	42	50	66	48	50	45
A	A	74	72	49	55	60	77	-	-	-
M	A	69	67	45	50	56	73	-	-	-
P	A	44	50	31	29	40	47	31	32	29

Table 1: **Classification using feature based representations.** Each row is a train / test pattern: **A**rt, **P**hoto, **M**ixed. Each column is an algorithm with feature, divided into groups: BoW [44, 4, 50, 15, 33, 46], Fisher Vectors [46]. Domain Adaption using GFK [29] in has two variants (PCA and LDA), also Subspace Alignment (SA) [22]. Each cell shows the; mean of 5 randomized trials. The standard deviation on any column never rises above 2%. Domain-Adaptation methods were tested only on cross-domain train/text patterns.

graph built by hierarchical agglomeration, filtered by Laplacian graph energy [53]. Tested using thirteen different classes in a heterogeneous database showed an accuracy (the diagonal of a confusion matrix) of above 85%. This suggests structural and spatial relations are important to cross-depiction; but the experiments are too limited to be conclusive and later tests on a larger dataset in [62] yields accuracies of around 20%, see Table 2. This suggests space and structure are important, but are insufficiently rich.

Given that proposition that features should not be limited by the statistics of any one domain (*e.g.* photograph, pencil drawing) we next considered simple shapes as features to label a graph. Using shapes are features was inspired by observing the great artists such a Picasso, who construct recognisable objects from circles, squares, and such like.

We first learnt shapes from image segmentation [61] using a fully unsupervised approach; unsupervised because we wanted to find out whether simple shapes exist in image segmentation independently of human bias. Our algorithm discovered simple shapes that can be named – circle, square, *etc.* These are seen in Figure 3.2.1. The same figure shows a scale-based hierarchical decomposition of an image with segments classified using these shapes, plus a "noise" category for segments that did not classify into any shape. A mean graph was used to connect shapes in each layer of the hierarchy, also in Figure 3.2.1. Edges also connect corresponding nodes between layers.

This was tested on a smaller image data base than in Sec. 3.1, and compared with dense SIFT [57] and structure only [2]. This representation maintains performance across domains – that is, it does not exhibit a fall-off when trained on one domain and tested on another, and all others do so far. Even so, a classification rate hovering around 60% cannot be regarded as satisfactory: we must turn to stronger models.

3.3 DPM, ADPM, and Multi-Label Graph

Deformable Parts Model ***DPM*** [18] is a well known object representation that takes spatial layout into account. It models an object with a star graph, *i.e.*, a root filter plus a set of parts. Given the location of the root and the relative location of n parts; $n = 8$ in our experiments. The score of the star model is the sum of responses of the root filter and parts filters, subtracting the displacement cost. Each node in a DPM is labelled with a HoG feature, learned from examples.

By analogy with domain adaptation, we considered

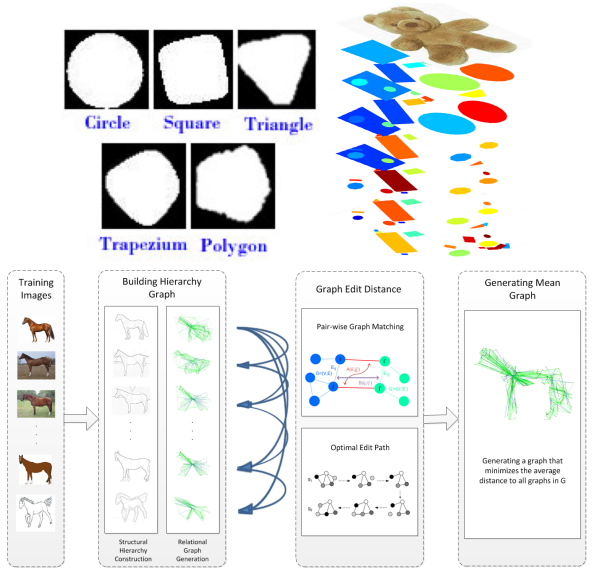

Figure 3: Top left: simple shapes learnt from segmentation without supervision. Top right: a hierarchy of shapes derived from an input image. Bottom: a mean graph learnt at head level in the hierarchy, with simple shapes labelling nodes. Edges also connect between layers.

case 1: Training	5p	5a
case 1: Testing	15p	15a
Dense SIFT	70%	59%
Structure Only	16%	19%
Proposed Method	61%	62%

case 2: Training	8p	10p	8a	10a
case2 : Testing	15a	15a	15p	15p
Dense SIFT	43%	47%	49%	51%
Structure Only	19%	23%	22%	25%
Proposed Method	63%	64%	64%	67%

case 3: Training	3a	5a	3p	5p
case 3: Testing	30m	30m	30m	30m
Dense SIFT []	46%	50%	50%	54%
Structure Only	13%	16%	14%	16%
Proposed Method	58%	61%	56%	61%

case 4: Training	6m	10m
case 4: Testing	30m	30m
Dense SIFT	60%	61%
Structure Only	21%	24%
Proposed Method	62%	65%

Table 2: **Classification using shape and structure**. From top to bottom, left to right: (a) single domain task, (b) single cross depiction task, and (c) single to mixture depiction task, (d) mixture cross depiction task. The character 'p' is 'photos', 'a' is 'art' and 'm' is 'mixture'. Dense SIFT was computed using [57], structure only follows [2].

the possibility of query expansion for DPM to obtain Adapted DPM (***ADPM)***. We first train a standard DPM model for each object category in the training set (*i.e.*, source domain) $\mathcal{S}$. We then apply the models on the test set (*i.e.*, target domain) $\mathcal{T}$. A confidence set $\mathcal{C} \subset \mathcal{T}$ is constructed from the test set for training expansion by picking images that match a particular VCM especially well:

$$\mathcal{C} = \{x \in \mathcal{T} | s_1(x) > \theta_1 \wedge s_1(x) - s_2(x) > \theta_2\} \quad (1)$$

with $s_1(x)$ the highest DPM score, and $s_2(x)$ the second highest score, and $\theta_1 \geq \theta_2$ are user-specified parameters to threshold the best score and margin respectively. We found $\theta_1 = -0.8$ and $\theta_2 = 0.1$ to be a good trade-off between minimising false positives (5%) and including appropriate number of expanded data (around 580 images in $\mathcal{C}$).

The fully connected multi-labelled graph (***MG***) model [59] is designed for the cross-depiction problem. It attempts to separate appearance features (contingent on the details of a particular depiction) from the information that characterises an object class without reference to any depiction. Unlike DPM, it comprises a fully connected weighted graph, and has multiple labels per node. Each graph has eight nodes. Weights on nodes can be interpreted as denoting the importance of a node to object class characterisation in a way that is independent of depiction. Weights on arcs are high if the distance between the connected pairs of parts varies little. These weights are learnt using a structural support vector machine [6]. In addition to the weights, each node carries 2 features labels. These are designed to characterise the appearance of parts in both photographs and artwork (see the Discussion 4 for a justification).

Table 3 compares the classification performance of DPM, ADPM and MG with the non-structure baseline FV. We can clearly see the benefit when considering the spacial information. Even so, the performance of standard DPM in 'train on photo, test on art' pattern significantly drops. However, this performance gap is shortened when the DPM model is re-learned on the expanded set, i.e., ADPM. It demonstrates that the expanded set does capture new information in the target domain and helps to refine the models according to the target domain. The MG alone maintains performance over all train/test patterns. The results suggest that structure and spatial layout is an essential information for recognising an object.

3.4 Deep Learning

Convolutional neural networks (CNN) [38] have yielded a significant performance boost on image classification. For classification, we follow Crowley and Zisserman [11], encoding images with CNN features, which are then used as input to learn a one-vs-all linear SVM classifier. The CNN parameters are pre-trained form the large ILSVR2013 dataset. We have included results from CNN in Table 2 because they compare so well with the space/structure aware methods. The pre-trained CNN achieved high performance when test on photos. Even so, CNNs exhibit the same fall in performance over the train-on-photo, test-on-art pattern that is seen in the feature based methods.

model		FV	DPM	ADPM	MG	CNN
train	test	SIFT	HOG	HOG	2×HOG	learnt
P	P	87	88	-	85	97
M	P	84	85	-	90	96
A	P	66	78	79	83	91
A	A	77	83	-	89	89
M	A	73	80	-	89	87
P	A	47	68	72	83	73

Table 3: **Classification using Space and Structure**. Each row is a train (30 image) / test (rest) pattern: **A**rt, **P**hoto, **M**ixed. Each column is an algorithm, Fisher Vectors [46], the best feature-only classifier, is repeated from Table 1. DPM[18] used a strong spatial layout model, ADPM is our domain adapted version. Multi-labelled graphs (MG)[59] has a stronger spatial model than DPM, and also has two labels at each node. We have include a deep learning CNN [11] too. Each cell shows the mean of 5 randomised trials. The standard deviation on any column never rises above 2%.

4 General Discussion

Across all experiments we see the same trend: a fall in performance in any case where art is included. This fall is most marked whenever photographs are used for training and artwork for testing, and is seen in all cases other than the Multi-Labelled Graph (MG) [59].

These observation need an explanation. Intuition suggests that the difference between the low-level images statistics of photographs and artwork is a cause. In particular, its is easy to imagine that the variation in low-level statistics across the gamut of all images is much wider than it is for any one depiction alone (photographs). This intuition is not ours alone, but is shared by others [11], but it remains untested.

A strong hypothesis is possible. Let $\mathbb{X}$ and $\mathbb{Y}$ be an object classes. Let $x_P \in \mathbb{X}$ be a photographic instance and x_A is artwork instance of class $\mathbb{X}$. Similarly $y_P, y_A \in \mathbb{Y}$ are a photograph and artwork of class Y. Denote the set of all x_p by X_P, meaning the 'photo visual object class $\mathbb{X}$', and likewise for X_A, Y_P, and Y_A. Suppose too there is a measure $d(,.)$ between each pair of elements in any set. The strong hypothesis is this: *The intra-class distance (same domain, different class) is expected to be less than the inter-class distance for (different domain, same class).* That is $d(x_P, x_A) > d(x_P, y_P)$, photographs are drawings of the same object are more different from each other than photographs of two different objects. Likewise, $d(x_P, x_A) > d(x_A, y_A)$, *etc.* To test this we used raw images Photo-Art-50 as raw input, each scaled to a square image of pixel width 256. We then mapped all the data into a 4 dimensional space using PCA over all the data (which captured most of the eigenenergy). We assumed a K-NN classifier, so that X_P is represented by the mean, likewise X_A. The measure, $d(.,.)$, is Euclidean distance. We found a fraction 0.67 of all statements of the form $d(x_P, x_A) > d(x_P, y_P)$ *etc.* to be true, which supports the stronger hypothesis. Figure 4 illustrates, showing that for all classes the different domains art/photograph tend to separate. This result explain our results above: a density fitted to photographic features alone is over-fitted because it fails to generalise to art-like features, and vice-versa. Wu *et al* [59] describe feature distributions using more than one centre, and are the most consistent of all descriptions over all recognition tasks on the Photo-Art-50 dataset.

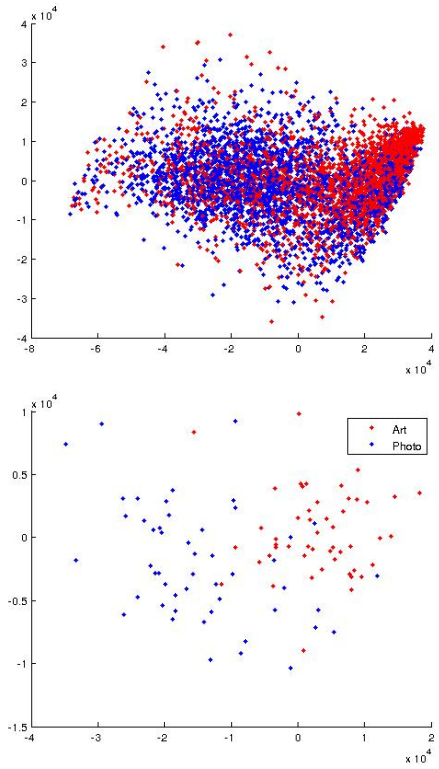

Figure 4: Above: each image in Photo-Art-50 plotted in an eigenspace spanning raw images, art in red, photos in blue. Below: The centre of each class in Photo-Art-50: red (art), blue(photo). The images and the cluster centres tend to form two groups: art/photo.

This wide variance in low-level statistics also helps explain the value of spatial information regarding object class identity. So far every method we have experimented that uses some kind of spatial information shows less fall away in the cross-depiction problem; this is true also of [12]. In this paper we see DPM outperform BoW, and the MG outperform DPM. This result is in line with (*e.g.*) Leordeanu *et al.* [42] who use the

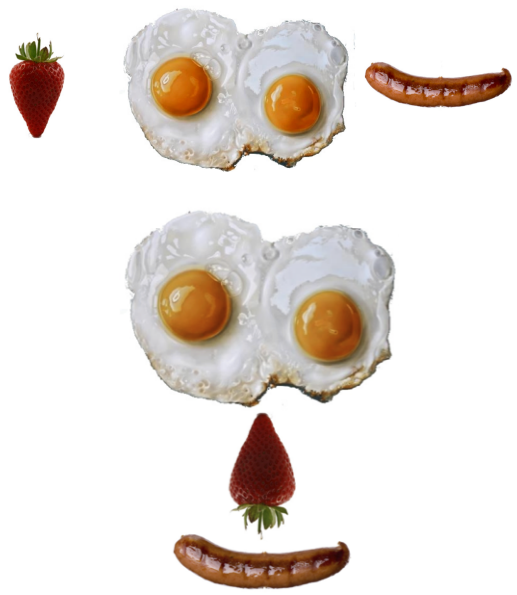

Figure 5: The presence of a face depends on spatial arrangement of parts: above, no face; below smiling face.

distance between low-level parts (edgelets) as a feature to characterise objects and achieve excellent detection results on the PASCAL dataset [17] of photographs; it may be effective too on Photo-Art-50, but this is to be proven.

This empirical data has anecdotal evidence too. The children's drawings in Figure 1 are clearly people, but have little in common with photographs of people, and not much in common with one another. Consider too Figure 5 in which the same parts form a face, or not, depending only on the spatial arrangement of the parts. Indeed, artwork from prehistory to the present day, whether produced by a professional or a child, no matter where in the world: the greater majority of it relies on spatial organisation for recognition. It is as if spatial organisation provides a major class, which is refined using features such as shape; but we have no direct evidence for this conjecture.

5 Cross Depiction Synthesis

Photorealistic image generation is common in Computer Graphics. Here we focus only on Non-Photorealisic Rendering (NPR) from photographs.

Structure, spatial layout, and shape are all important characteristics in identifying objects regardless of depiction. Equally, they can be used to generate artwork directly from photographs. Consider Figure 6; it shows a photograph of a bird feeding its young. The photograph has been segmented, and the segments classified into one of a few qualitative shapes (square, circle, triangle, ...). In the most extreme case just one class (circle) is used. See [54] for details of the computer graphics algorithm.

It is true that as the degree of abstraction grows the original interpretation of the image becomes harder to

Figure 6: Shape abstraction for Automated Art.

maintain; but given too the degree of abstraction in children's drawings, the conclusion that both the quality and quantity of abstraction is important for recognition. In this case the aim was only to produce a "pretty" image that bears some resemble to the original. However, simple qualitative shapes of the kind used here can be learned directly from segmentation, as are sufficient to classify scene type (indoor, outdoor, city ...) at close to state-of-the-art rates [60].

Shape is not the only form of abstraction useful to the production of art, structure can be used too. Figure 7 shows examples of computer generated art based on rendering structure. The analysis used to obtain the structure is identical to that used by [2] to classify objects based on weighted graphs alone. In this case the arcs of a graph have been visualised in a non-photorealistic manner, and the shape of parts at nodes have been classified into a qualitative shape; see [32] for details, which specified the shapes learnt from segmentation by [60].

6 Conclusion

It is clear that the same sorts of representations that support abstract image synthesis also support image classification. It seems that synthesis and classification are indeed related, as intuition would have us believe.

The cross-depiction problem pushes at the foundations of computer vision, because it brings in sharp focus the question of how to describe object classes. Given the fact that the same kind of representations are used both for abstract rendering and for recognition, the conclusion that there is a strong relation between the two is hard to escape. The relation between the cross depiction problem and image generation is given (strong) anecdotal support by the observation that people draw a mix of what they know and what they see. We can see this in the art of children, and by the fact that when draughting was considered important, by Art Schools, the tutors had to train students to draw what they see rather than what they know – that is one of the main purposes of life-drawing classes.

Our experimental results show that recognition algorithms premised directly on *appearance* suffer a fall in performance within the cross-depiction problem; probably because they tacitly assume limited variance of low-level statistics. Rather, they suggest that struc-

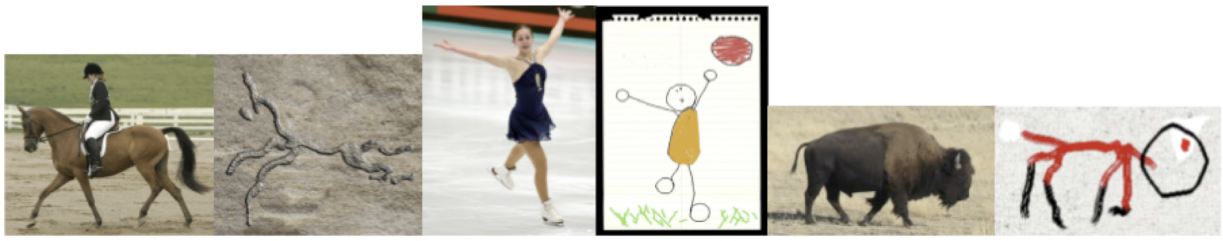

Figure 7: Structure and Shape combine to make art in the style of (left to right) petroglyphs, child art, Joan Miro.

ture, spatial layout, and shape are all important characteristics in identifying objects regardless of depiction.

For example, DPM out-performs BoW-HOG, even though both use the same low level features; the MG – with a stronger spatial model – out-performs DPM. This is because, possibly, structure and spatial layout capture the essential form of an object class, with specific appearance relegated to the level of detail. In other words, structure and space are more salient to robust identification that appearance. Indeed all algorithms we have tested show a significant fall compared to their own peak in performance, when trained on photographs and tested on art; this includes the deep learning methods we have used. The single exception is ([59]), which explicitly models a strong structure, and explains appearance details using multiple labels on each node (multiple labels to account for both art and photographic appearance).

The relative importance and the interaction between the descriptors we have identified as important remains an open problem, and does the possibility of other descriptive terms has not been eliminated. A zebra and a horse look largely identical, except for texture.

Deep learning performs very well on classification over Photo-Art-50, but it does exhibit a fall in performance when trained on photographs and tested on art – only the multi-labelled graph [59] and the (lesser performing) graph-with-shapes [62] do not. Also, we have found that when presented with the problem of people detection in a much larger database CNN methods do not rise above a detection rate of 40%. These results make it difficult to conclude that deep learning is a solution to the cross-depiction problem; quite possibly it too suffers from lack of spatial awareness.

In summary: the cross-depiction problem pushes the envelope of computer vision research. It offers significant challenges, which if solved will support new applications in computer graphics and other areas. Modelling visual classes using structure and spatial relations seems to offer a useful way forward; the role of deep learning in the problem is yet to be fully proven in comparison to its own performance in other tasks and when compared to human ability in this difficult challenge.

References

[1] Yali Amit and Alain Trouvé. Pop: Patchwork of parts models for object recognition. *International Journal of Computer Vision*, 75(2):267–282, 2007.

[2] Xiao Bai, Yi-Zhe Song, and Peter Hall. Learning invariant structure for object identification by using graph methods. *Computer Vision and Image Understanding*, 115(7):1023–1031, 2011.

[3] BBC. Your paintings dataset. http://www.bbc.co.uk/arts/yourpaintings/.

[4] Alexander C. Berg and Jitendra Malik. Geometric blur for template matching. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.

[5] K. Chatfield, J. Philbin, and A. Zisserman. Efficient retrieval of deformable shape classes using local self-similarities. In *Workshop on Non-rigid Shape Analysis and Deformable Image Alignment, ICCV*, 2009.

[6] M. Cho, K. Alahari, and J. Ponce. Learning graphs to match. In *ICCV*.

[7] John P Collomosse, Graham McNeill, and Yu Qian. Storyboard sketches for content based video retrieval. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 245–252. IEEE, 2009.

[8] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[9] James Coughlan, Alan Yuille, Camper English, and Dan Snow. Efficient deformable template detection and localization without user initialization. *Computer Vision and Image Understanding*, 78(3):303–319, 2000.

[10] David Crandall, Pedro Felzenszwalb, and Daniel Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 10–17. IEEE, 2005.

[11] E.J. Crowley and A. Zisserman. In search of art. In *ECCV Workshop: VisArt*, 2014.

[12] Elliot Crowley and Andrew Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[13] Elliot J Crowley and Andrew Zisserman. Of gods and goats: Weakly supervised learning of figurative art. *learning*, 8:14, 2013.

[14] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[15] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, pages 886–893, 2005.

[16] Gal Elidan, Geremy Heitz, and Daphne Koller. Learning object shape: From drawings to images. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2064–2071. IEEE, 2006.

[17] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[18] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[19] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[20] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[21] Robert Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–264. IEEE, 2003.

[22] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.

[23] Vittorio Ferrari, Loic Fevrier, Frederic Jurie, and Cordelia Schmid. Groups of adjacent contour segments for object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(1):36–51, 2008.

[24] Vittorio Ferrari, Frederic Jurie, and Cordelia Schmid. From images to shape models for object detection. *IJCV*, 2010.

[25] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.

[26] S. Ginosar, D. Haas, T. Brown, and J. Malik. Detecting people in cubist art. In *ECCV Workshop: VisArt*, 2014.

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[28] Boqing Gong, Kristen Grauman, and Fei Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, pages 222–230, 2013.

[29] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012.

[30] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *IEEE International Conference on Computer Vision*, volume 0, pages 999–1006, 2011.

[31] Chunhui Gu, Joseph J Lim, Pablo Arbelaez, and Jitendra Malik. Recognition using regions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1030–1037. IEEE, 2009.

[32] Peter Hall and Yi-Zhe Song. Simple art as abstractions of photographs. In *Proceedings of the Symposium on Computational Aesthetics*, pages 77–85. ACM, 2013.

[33] Rui Hu, Mark Barnard, and John P. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, pages 1025–1028, 2010.

[34] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117(7):790–806, 2013.

[35] Rui Hu, Stuart James, Tinghuai Wang, and John Collomosse. Markov random fields for sketch based video retrieval. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 279–286. ACM, 2013.

[36] Wei Jia and Stephen J McKenna. Classifying textile designs using bags of shapes. In *ICPR*, pages 294–297, 2010.

[37] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[39] Jan Eric Kyprianidis, John Collomosse, Tinghuai Wang, and Tobias Isenberg. State of the" art&# x201d;: A taxonomy of artistic stylization techniques for images and video. *Visualization and Computer Graphics, IEEE Transactions on*, 19(5):866–885, 2013.

[40] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE, 2006.

[41] Bastian Leibe, Aleš Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *International journal of computer vision*, 77(1-3):259–289, 2008.

[42] M. Leordeanu and M. Herbert andR. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *CVPR*, 2007.

[43] Yi Li, Yi-Zhe Song, and Shaogang Gong. Sketch recognition by ensemble matching of structured features. In *In British Machine Vision Conference (BMVC)*. Citeseer, 2013.

[44] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, 2004.

[45] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang 0001. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[46] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 143–156, 2010.

[47] Hillel Rom and Gerard Medioni. Hierarchical decomposition and axial shape description. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(10):973–981, 1993.

[48] Olga Russakovsky, Yuanqing Lin, Kai Yu, and Li Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

[49] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision*, pages 213–226, 2010.

[50] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[51] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics (TOG) (Proceedings of ACM SIGGRAPH ASIA)*, 30(6), 2011.

[52] Kaleem Siddiqi, Ali Shokoufandeh, Sven J Dickinson, and Steven W Zucker. Shock graphs and shape matching. *International Journal of Computer Vision*, 35(1):13–32, 1999.

[53] Yi-Zhe Song, Pablo Arbelaez, Peter Hall, Chuan Li, and Anupriya Balikai. Finding semantic structures in image hierarchies using laplacian graph energy. In *Computer Vision–ECCV 2010*, pages 694–707. Springer, 2010.

[54] Yi-Zhe Song, David Pickup, Chuan Li, Paul Rosin, and Peter Hall. Abstract art by shape classification. *Visualization and Computer Graphics, IEEE Transactions on*, 19(8):1252–1263, 2013.

[55] Hari Sundar, Deborah Silver, Nikhil Gagvani, and S Dickinson. Skeleton based shape matching and retrieval. In *Shape Modeling International, 2003*, pages 130–139. IEEE, 2003.

[56] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.

[57] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[58] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[59] Qi Wu, Hongping Cai, and Peter Hall. Learning graphs to model visual objects across different depictive styles. In *Computer Vision–ECCV 2014*, pages 313–328. Springer, 2014.

[60] Qi Wu and Peter Hall. Prime shapes in natural images. In *BMVC*, pages 1–12, 2012.

[61] Qi Wu and Peter Hall. Prime shapes in natural images. In *Proceedings of the British Machine Vision Conference*, pages 45.1–45.12. BMVA Press, 2012.

[62] Qi Wu and Peter Hall. Modelling visual objects invariant to depictive style. In *BMVC*, 2013.